

# O-MaMa: Learning Object Mask Matching between Egocentric and Exocentric Views

Lorenzo Mur-Labadia\* Maria Santos-Villafranca\* Jesus Bermudez-Cameo

Alejandro Perez-Yus Ruben Martinez-Cantin Jose J. Guerrero

I3A - University of Zaragoza

## Abstract

Understanding the world from multiple perspectives is essential for intelligent systems operating together, where segmenting common objects across different views remains an open problem. We introduce a new approach that re-defines cross-image segmentation by treating it as a mask matching task. Our method consists of: (1) A Mask-Context Encoder that pools dense DINOv2 semantic features to obtain discriminative object-level representations from FastSAM mask candidates, (2) an Ego $\leftrightarrow$ Exo Cross-Attention that fuses multi-perspective observations, (3) a Mask Matching contrastive loss that aligns cross-view features in a shared latent space, and (4) a Hard Negative Adjacent Mining strategy to encourage the model to better differentiate between nearby objects. O-MaMa achieves the state of the art in the Ego-Exo4D Correspondences benchmark, obtaining relative gains of +22 % and +76 % in the Ego2Exo and Exo2Ego IoU against the official challenge baselines, and a +13 % and +6 % compared with the SOTA with 1 % of the training parameters.

## 1. Introduction

Nowadays, intelligent agents need to collaborate while performing cooperative tasks. This includes applications like multi-robot manipulation [29, 64], augmented reality assistants [70, 74], and human-robot collaboration [7, 13]. Perception plays a crucial role in this scenario, but each agent typically has access only to its own sensors or cameras, and each one perceives the environment from a different perspective. Consequently, understanding object corre-

\*Equal contribution.

Corresponding authors: [lmur@unizar.es](mailto:lmur@unizar.es),

[m.santos@unizar.es](mailto:m.santos@unizar.es)

Code: <https://github.com/Maria-SanVil/O-MaMa>

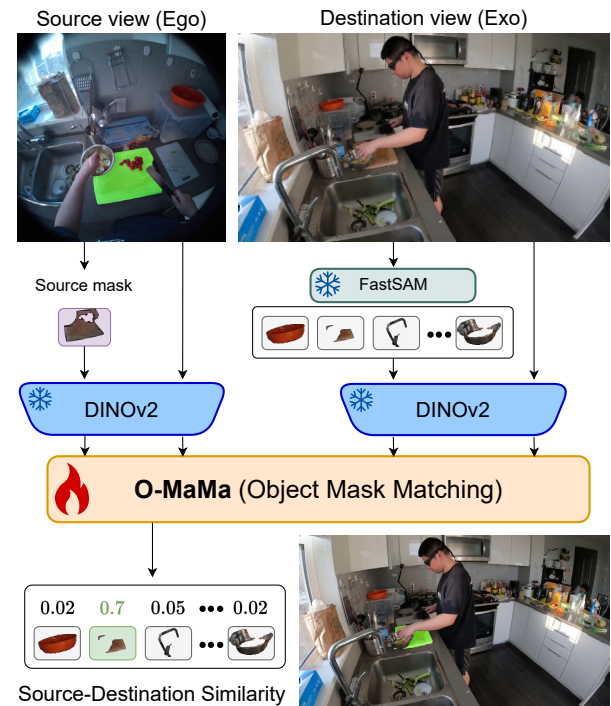


Figure 1. Overview of our proposed Object Mask Matching (O-MaMa). Instead of attempting the complex cross-view segmentation task, we obtain a set of mask candidates in the destination view using FastSAM. Through contrastive learning, we select the mask candidate that best matches the source mask.

spondences between egocentric (first-person) and exocentric (third-person) views is essential to align multi-agent perception and establish a shared basis for interaction. Despite the advances in segmentation [40, 43, 47, 77] and object detection [5, 25] from single images, cross-view segmentation between egocentric and exocentric perspectives remains an open challenge.

In this paper, we address the Ego-Exo Correspondences task, where the goal is to predict an object’s mask in one perspective given a query mask from the other. Unlike traditional segmentation problems, this task introduces additional challenges, including drastic viewpoint transformations, scale variations, occlusions, and domain shifts due to differences in camera optics and imaging conditions. While the exocentric view captures both the full environment and the person’s body, it contains objects at multiple scales. Conversely, the egocentric view offers rich details on hand-object interactions, but it is highly dynamic, suffering from motion blur and frequent occlusions due to the ongoing interactions. These challenges make establishing precise object-level correspondences across views particularly difficult, requiring fine-grained segmentation and strong cross-view semantic reasoning.

We propose simplifying the complex cross-image segmentation task by reformulating it as Object Mask Matching (O-MaMa) across ego and exo views, leveraging the excellent zero-shot segmentation capabilities of Segment Anything Models (SAM) [33]. An overview of our method is shown in Fig. 1. First, we extract a set of object mask candidates in the destination view using FastSAM [87]. To obtain mask descriptors, we encode each object mask with a *Mask-Context Encoder*, which pools dense DINOv2 [55] semantic features from both object masks and their extended bounding boxes, combining discriminative object features with contextual information. In cluttered scenes, nearby objects often share similar context while containing distinct object descriptors. Therefore, we introduce a *Hard Negative Adjacent Mining* strategy that selects neighboring object candidates to encourage the model to better differentiate between nearby objects. Next, we incorporate cross-view global features using a novel *Ego↔Exo Cross Attention* mechanism. Finally, we train the model with a *Mask Matching Contrastive Loss*, which selects the best mask candidate in the destination view by learning a cross-view feature alignment that captures both global scene context and fine-grained object features.

Our method is both simple and effective, achieving state of the art in the Ego-Exo4D Correspondences benchmark [24]. We obtain 45.8 (Ego2Exo) and 48.6 (Exo2Ego) IoU in the test v2 split, which represents a relative gain of +22.1 % and +76.4% against the official challenge baselines [24], respectively. Moreover, in the validation v1 set, O-MaMa scores 50.1 and 54.2 IoU in the Ego2Exo and Exo2Ego scenarios, showing a +13.1 % and +6.5 % against the previous SOTA model [19] using only 1 % of training parameters.

## 2. Related Works

**Ego-Exo understanding.** Exocentric (third-person) vision has been extensively studied in action recognition [26, 36], segmentation [31, 71, 76] and tracking [12, 75]. Al-

ternatively, egocentric (first-person) vision offers a unique viewpoint for capturing human behavior. The arrival of large-scale datasets [14, 23] has driven progress in action recognition [1, 58], action anticipation [20, 51], affordance segmentation [49, 52], and episodic memory [3, 46]. Since egocentric and exocentric views provide complementary visual cues of the scene, combining both perspectives has emerged as a promising direction for learning generalizable view-invariant representations. Some works [17, 39] improve model training in one perspective by leveraging data properties from the other view. Other works explore the benefits of learning cross-view invariant features for action recognition [60, 80, 83], affordance segmentation [35], activity progression [16], and temporal action segmentation [57]. In contrast, we propose learning view invariant features at object level to match masks across synchronized views, which requires fine-grained pixel-level predictions.

**Learning correspondences.** Traditionally, detector-based local feature matching methods first detect keypoints in an image and then extract descriptors to establish correspondences following hand-crafted [4, 45, 62] or learning [15, 82] approaches, using nearest neighbor search or an attentional graph neural network [63] to find matches between the extracted interest points. More recently, detector-free methods [18, 28, 67] have gained attention for their ability to directly obtain dense feature matches without an explicit keypoint detection. For instance, RoMa [18] achieves robust feature matching under extreme changes in scale, viewpoint and illumination. While these methods establish correspondences between images of the same scene or object instances, semantic correspondence approaches [11, 42, 84] find dense matches between semantically similar images. For example, Zhang *et al.* [84] fuses DINOv2, that provides sparse but accurate matches, with stable diffusion features, which contain high-quality spatial information. Here, we propose leveraging powerful semantic features from DINOv2 as well, but in a different context. Instead of keypoints or semantic parts, we establish correspondences between entire object masks across different perspectives.

**Segmentation Models.** Traditional approaches are categorized into semantic segmentation [43, 77], instance segmentation [5, 25, 50], panoptic segmentation [8, 9, 32, 48], and video object segmentation [56, 69, 81]. SAM [33] introduced the promptable segmentation task [30, 38, 78, 87], where fine-grained segmentation masks are generated from a spatial prompt (*i.e.*, a point, or a bounding box). Recently, several works [34, 86, 89] have leveraged the input token flexibility of Large Language Models (LLMs) to support diverse input conditions and output formats.

Despite the significant advances in single-image segmentation, few works address object segmentation across multiple images. For instance, semi-supervised video ob-

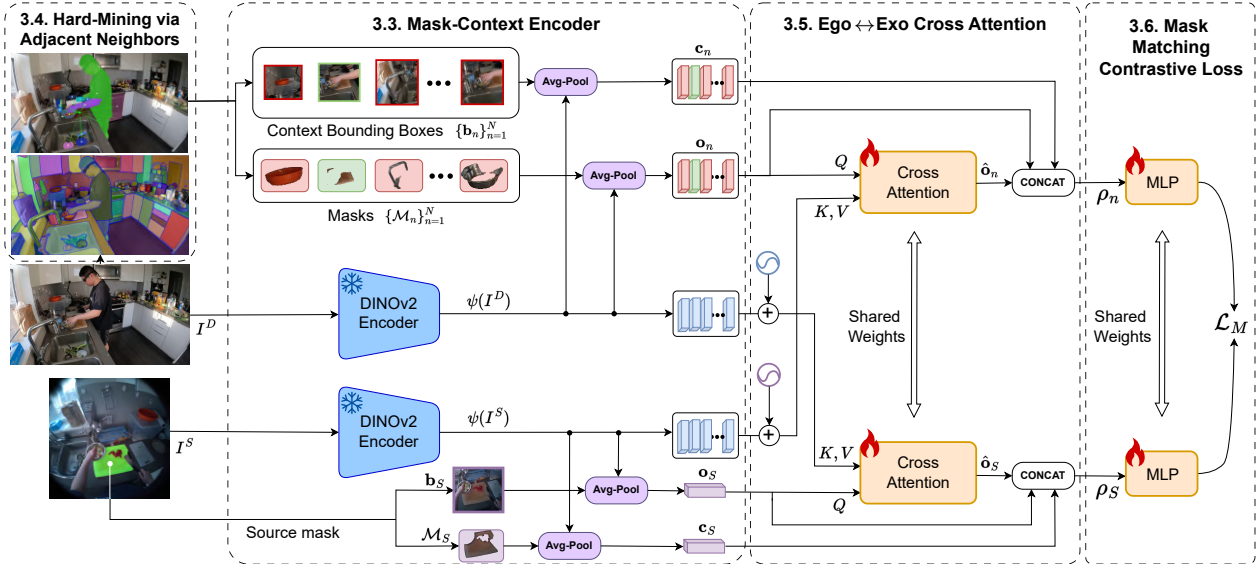


Figure 2. **O-MaMa architecture.** In the destination view, we generate a set of mask candidates with FastSAM. We extract descriptors on both source and destination masks by pooling dense DINOv2 features, and we aggregate global cross-view features with respective cross-attention mechanisms. We learn view-invariant features in a latent space via contrastive learning, and we select the most similar mask embedding to obtain the corresponding mask.

ject segmentation [56] methods segment an object along the video given its mask in the initial frame [6, 22, 37, 53, 79]. Similarly, image co-segmentation models aim to identify common objects across different images [61, 65, 68]. Adapting LLMs to the cross-view segmentation, Object-Relator [19] fine-tuned PSALM [86] with a dedicated module that enforces view-invariant embeddings through a self-supervised alignment. In this work, we propose to reformulate the cross-view segmentation as a mask matching task, incorporating FastSAM [87] in our pipeline to leverage its fine-grained, zero-shot segmentation capabilities; showing that we can achieve state-of-the-art results while significantly reducing the number of trainable parameters.

### 3. Methods

#### 3.1. Task Formulation

Given a pair of images from two different views (egocentric and exocentric) and a query object mask  $\mathcal{M}_S$  in the source view  $I^S$ , the objective of the Ego-Exo Correspondences task is to predict the corresponding object mask in the destination view  $I^D$ . In the Ego2Exo task, the source view corresponds to the egocentric image, and the destination view is the exocentric image, and the opposite happens in the Exo2Ego setting. The task is restricted to use only visual information as input for the model, excluding semantic labels, object names, or camera pose information.

Additionally, the task involves significant viewpoint variation and challenges due to the inherent characteristics

of each viewpoint: egocentric views suffer camera motion and occlusions due to the ongoing interactions, while the exocentric perspective contains objects at multiple scales distributed along all the scene. See Fig. 3 for some examples of these limitations in the dataset. All these factors require combining both a very fine-grained segmentation capability and a global cross-view understanding to effectively locate the corresponding objects across views.

#### 3.2. Method overview

We reformulate the challenging cross-view segmentation task as a cross-view Object Mask Matching (O-MaMa) problem. This novel approach exploits the high-quality zero-shot segmentation capabilities exhibited by SAM [33, 87] to simplify the segmentation problem. Fig. 2 shows the architecture of our method. The goal of O-MaMa is to select the object mask candidate in the destination view that matches the source mask best. We first extract a set of object mask proposals in the destination view using FastSAM [87], from which we compute object-level and contextual descriptors with the *Mask-Context Encoder* (Section 3.3). We also include cross-view global features through a novel *Ego↔Exo Cross Attention* mechanism (Section 3.5). We adopt a novel *Mask Matching Contrastive Loss* (Section 3.6) to learn view-invariant features by minimizing the distance between paired samples and maximizing the distance between negative (unpaired) samples in a shared latent space. During training, we enforce the model to learn more robust and discriminative object descriptors with a



Figure 3. **Examples of complex scenarios.** The object may appear on the edges, be partially occluded or be extremely small.

*Hard-Mining via Adjacent Neighbors* (Section 3.4).

### 3.3. Mask-Context Encoder

We first generate a set of dense mask proposals in the destination view using FastSAM [87], which segments intuitive regions of the scene, such as entire objects, object parts or surfaces, with comparable quality to SAM [33], while being  $50\times$  faster. Specifically, from the destination image,  $I^D$ , we generate  $N$  mask candidates  $\{\mathcal{M}_n\}_{n=1}^N$ .

Then, we compute a descriptor of each mask segment. We leverage DINOv2 [55], a self-supervised learning model, due to its high-level semantics, object decomposition capabilities and dense feature localization properties. We extract a local object descriptor  $\mathbf{o}_n = \text{Avg-Pool}(\mathcal{M}_n, \psi(I^D))$  by pooling the corresponding object mask from the DINOv2 feature map of the destination image, denoted as  $\psi(I^D)$ . Some studies suggest [2, 21, 27] that humans leverage visual contextual associations among objects to represent scenes. Inspired by this, we extract a context descriptor  $\mathbf{c}_n = \text{Avg-Pool}(\mathbf{b}_n, \psi(I^D))$  by pooling features from an extended bounding box  $\mathbf{B}$  around each mask. In both cases, we upsample  $\times 4$  the DINOv2 feature map size in order to retain feature’s regions fine granularity [66]. Similarly, we extract object  $\mathbf{o}_S$  and context embeddings  $\mathbf{c}_S$  of the source mask  $\mathcal{M}_S$  in the other view.

### 3.4. Hard-Mining via Adjacent Neighbors

While the object embedding  $\mathbf{o}_n$  contains very discriminative object features, the context embedding  $\mathbf{c}_n$  incorporates surrounding information to help localizing the object in the other view, but this surrounding context also introduces ambiguity in cluttered environments, where nearby objects share a similar context. To address this, we introduce a hard-negative mining strategy based on adjacent neighbors, encouraging the model to disambiguate between nearby but distinct objects with similar context. In the destination view, we construct a graph of mask segments based on the pixel centers of each mask using the Delaunay Triangulation, as Fig. 4 shows. This results in a binary adjacency matrix  $\mathcal{A} \in \{0, 1\}^{N \times N}$ , defining the connectivity between segments. We define  $\mathcal{N}(\mathbf{o}_n)$  to the set of neighbors of object  $\mathbf{o}_n$ . Then, we take the second order neighbor set  $\mathcal{N}^2(\mathbf{o}_n) = \{\mathcal{N}(\mathbf{o}_j) \mid \forall \mathbf{o}_j \in \mathcal{N}(\mathbf{o}_n)\}$ . Finally, we consider



Figure 4. **Hard Negatives mining examples.** We visualize  $2^{nd}$  order adjacent neighbors both in ego (left) and exo (right) scenarios.

the joint set of first and second order neighbors as the hard negative candidates  $\mathcal{O}_n^- = \{\mathcal{N}(\mathbf{o}_n) \cup \mathcal{N}^2(\mathbf{o}_n)\}$ .

### 3.5. Ego↔Exo Cross Attention

Although the mask context embedding incorporates surrounding contextual information, it lacks a global representation across views. Therefore, we introduce a *Ego↔Exo Cross Attention* mechanism, which enhances the object embedding by extracting its corresponding semantic features in the other view. Specifically, we compute a cross attention operation [73] between the candidate object masks  $\mathbf{o}_n$  and the source image feature map  $\psi(I^S)$ .

$$\hat{\mathbf{o}}_n = \text{Softmax} \left( \frac{\mathbf{o}_n W_Q \cdot (\psi(I^S) W_K)^\top}{\sqrt{d}} \right) \cdot \psi(I^S) W_V$$

We compute query vectors from the candidate object descriptors  $\mathbf{o}_n$  with a linear projection  $W_Q$ , while key and value vectors represent the overall source image features  $\psi(I^S)$  using the  $W_K$  and  $W_V$  linear layers. Before the cross-attention operation, we incorporate a learnable positional embedding to encode the spatial location of the patch tokens and a standard Layer Norm operation. Intuitively, the cross-view embedding of the object-mask candidates  $\hat{\mathbf{o}}_n$  captures how each potential mask candidate is represented in the source view. Similarly, we compute the cross-view embedding of the source mask  $\hat{\mathbf{o}}_S$  using the source mask descriptor  $\mathbf{o}_S$  for the queries and the overall destination image features  $\psi(I^D)$  for the keys and values.

### 3.6. Mask Matching Contrastive Loss

The final descriptor  $\rho_n$  is obtained from the  $n$ -th candidate mask by concatenating its refined cross-view embedding  $\hat{\mathbf{o}}_n$ , the context embedding  $\mathbf{c}_n$  and the object embedding  $\mathbf{o}_n$ ; while the final descriptor  $\rho_S$  of the source mask is the result of concatenating  $\hat{\mathbf{o}}_S$ ,  $\mathbf{c}_S$  and  $\mathbf{o}_S$ , respectively. Then, a shallow multi-layer perceptron  $f_\theta(x) \in \mathbb{R}^{d_f}$  maps the cross-view embeddings to a latent feature representation, where  $d_f$  is the dimension of the common feature space. This mapping ensures that both egocentric and exocentric masks are embedded within a shared latent space, enabling a cross-view comparison.

Our contrastive loss is based on InfoNCE [54]. We select a batch  $\mathcal{B}$  of  $|\mathcal{B}|$  elements, one positive and  $|\mathcal{B}| - 1$  negatives from the list of closest neighbors around the target object in the other view  $\mathcal{O}_n^-$  as defined in Section 3.4. If the number of neighbors is greater than the negative batch size,  $|\mathcal{O}_n^-| > |\mathcal{B}|$  we randomly select a subset of  $|\mathcal{B}|$  elements from the neighbor set  $\mathcal{O}_n^-$ . If  $|\mathcal{O}_n^-| < |\mathcal{B}|$  we also include masks from random objects from the rest of the image. If the number of segmented objects is less than the neighbor set  $\mathcal{O}_n^-$ , the objects are duplicated in order to fill the negative batch size  $|\mathcal{B}|$ . Finally, we apply the pairwise cosine similarity  $\text{sim}(\cdot, \cdot)$  between the source mask embedding  $f_\theta(\rho_S)$  and the batch of  $|\mathcal{B}|$  mask candidates embeddings  $\{f_\theta(\rho_n)\}_{n=1}^{|\mathcal{B}|}$  for computing the training loss:

$$\mathcal{L}_M(\rho^+, \rho_S) = -\log \frac{\exp(\text{sim}(f_\theta(\rho^+), f_\theta(\rho_S))/\tau)}{\sum_{n=1}^{|\mathcal{B}|} \exp(\text{sim}(f_\theta(\rho_n), f_\theta(\rho_S))/\tau)}$$

where  $f_\theta(\rho^+)$  is the positive element in the batch. This *mask matching contrastive loss* aligns the corresponding cross-view object embeddings while it separates the remaining object candidates in the shared feature space.

### 3.7. Inference

At inference, we choose the object candidate whose embedding is closest to the source object in the latent space.

$$\mathcal{M}_{n^*}, \text{ where } n^* = \arg \max(\text{sim}[f_\theta(\rho_n), f_\theta(\rho_S)])$$

Intuitively, the similarity score should be low for objects that do not share context or appearance.

## 4. Experiments

### 4.1. Experimental Setup

**Training dataset.** We use the novel Ego-Exo4D [24] dataset for our experiments. Ego-Exo4D is a massive-scale multi-modal video dataset containing synchronized egocentric and exocentric recordings of human activities. Specifically, we consider the Ego-Exo4D Correspondences set, which includes 1.8M synchronized object masks annotated at 1 FPS, covering 5.6K objects across 1335 unique videos and six different activities (*cooking, bike repair, health, music, basketball, and soccer*).

**Evaluation.** Following the official Ego-Exo4D Correspondences benchmark [24], we adopt the Intersection over Union (IoU) as the primary evaluation metric. We also report the Contour Accuracy (Cont.A) [56] to measure the similarity between predicted and ground-truth mask contours after translation, and the Location Error (Loc.E), which quantifies the normalized distance between the predicted and ground-truth mask centroids.

**Implementation details.** We employ FastSAM [87] with default hyper-parameters (0.9 IoU, 0.4 confidence score)

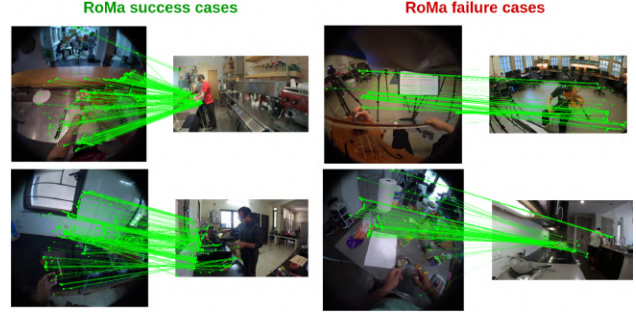


Figure 5. **RoMa [18] success and failure cases.** The extreme view variance makes that, even SOTA methods in geometry matching like RoMa [18], fail in extracting matches.

to generate dense candidate masks in the destination view. FastSAM achieves performance comparable to SAM [33] while being  $50\times$  faster and utilizing just 68M parameters. For feature extraction, we use a DINOv2 [55] ViT-B/14 model, which consists of 86M parameters. Our model is trained with the AdamW optimizer [44] and an initial learning rate of  $8 \cdot 10^{-5}$  with cosine annealing scheduling. We use a batch size of 24 image pairs, sampling 32 masks candidates in each destination image during training. We conduct our experiments on two NVIDIA GeForce RTX 4090.

### 4.2. Baseline Models

We compare our approach against the following baselines<sup>1</sup>:

- **XSegTx and XView-XMem** are the official baselines [24]. XSegTx adapts an image co-segmentation model [65], extended with a cross-view temporal memory [10].
- **CMX [85]** is a transformer-based segmentation model that fuses two modalities. We concatenate the query mask with the source image, and we adapt the decoder to predict the mask.
- **PSALM [86]** combines a LLM with Mask2Former [9] to perform zero-shot segmentation. **ObjectRelator [19]** trains PSALM with specialized cross-view modules.
- **k-Nearest Neighbors (k-NN)**. This is a naïve version of our approach. We extract descriptors of the generated mask candidates in the destination view, and we select the most similar to the query mask in the source view.
- **Geometry Methods [18]**. We restrict the k-NN search to only the masks that satisfy the epipolar line restriction, in order to evaluate the geometrical constraints of traditional geometrical matching methods. We tried LightGlue [41] using either SuperGlue [63] (43.8 % success rate<sup>2</sup>), SIFT [45] (42 % success rate), DISK [72] (31.2 % success rate) or ALIKED [88] (40.2 % success rate), and compared them using RoMa [18] (67.6 % success rate).

<sup>1</sup> See the supplementary material for implementation details

<sup>2</sup> Rate of success cases (see Fig. 5) when extracting matches in pairs of images. Refer to the supplementary material for more details.

Method	Ego2Exo			Exo2Ego			Total-IoU $\uparrow$	Num. Param. (M)	
	IoU $\uparrow$	Loc.E $\downarrow$	Cont.A $\uparrow$	IoU $\uparrow$	Loc.E $\downarrow$	Cont.A $\uparrow$		Total	Train
PSALM [86] (Zero-shot)	7.4	0.266	0.121	2.1	0.294	0.058	4.8	1587.1	0
CMX [85]	6.8	0.110	0.137	12.0	0.166	0.177	9.4	138.0	17.3
XSegTx [24]	18.9	0.070	0.386	27.1	0.104	0.358	23.0	12.1	3.6
XMem [24]	19.3	0.151	0.262	16.6	0.160	0.240	18.0	62.2	62.2
XMem + XSegTx [24]	34.9	0.038	0.559	25.0	0.117	0.237	30.0	75.6	67.1
Ours (k-NN baseline)	31.9	0.195	0.414	30.9	0.127	0.373	31.4	154.0	0
Ours (O-MaMa)	<b>42.6</b>	<b>0.033</b>	<b>0.590</b>	<b>44.1</b>	<b>0.082</b>	<b>0.524</b>	<b>43.4</b>	165.6	11.6

Table 1. Results on the Ego-Exo4D Correspondences v2 test split.

Method	Ego2Exo IoU $\uparrow$	Exo2Ego IoU $\uparrow$	Total IoU $\uparrow$	Total Param.(M)	Train Param.(M)
XSegTx	6.2	30.2	18.2	12.1	3.6
XMem	17.2	20.7	19.0	62.2	62.2
XMem + XSegTx	36.9	36.1	36.5	75.6	67.1
PSALM (zero-shot)	7.9	9.6	8.8	1587.1	0
PSALM (fine-tuned)	41.3	44.1	42.7	1587.1	1587.1
ObjectRelator	44.3	50.9	47.6	1587.3	1587.3
Ours (k-NN baseline)	40.5	40.6	40.6	154.0	0
Ours (O-MaMa)	<b>50.1</b>	<b>54.2</b>	<b>52.1</b>	165.6	11.6

Table 2. Ego-Exo4D Correspondences v1 val split results.

Due to its more view variance robustness, we select this last method to obtain the fundamental matrix and transfer the mask centroid in the source view to its epipolar line in the destination view, and we discard those candidate masks further than a certain threshold to the epipolar line.

### 4.3. Comparison with the State of the Art

Tab. 1 presents results on the EgoExo4D Correspondences v2 test set, demonstrating the our approach’s effectiveness. Even our simplest version, the k-NN baseline, already surpasses the official XMem+XSegTx, achieving 31.9 IoU in Ego2Exo and 30.9 IoU in Exo2Ego tasks. Our full method, O-MaMa, further improves performance, reaching 42.6 Ego2Exo and 44.1 Exo2Ego IoU, representing considerable relative gains<sup>3</sup> of up to +22.1 % and +76.4 % over XMem+XSegTx. The improvement is consistent in the other metrics, where O-MaMa obtains 0.033 Loc.E, 0.590 Cont.A in the Ego2Exo task and 0.082 Loc.E and 0.524 Cont.A in the Exo2Ego task.

As ObjectRelator [19] reports only results on the outdated EgoExo4D Correspondences v1 validation split, we also show results in this split in Tab. 2. O-MaMa also achieves the best performance, obtaining a 50.1 Ego2Exo IoU and 54.2 Exo2Ego IoU, while requiring only 1 % of the trainable parameters compared to [19]. The potential of our approach is further demonstrated by comparing our k-NN baseline with PSALM [86] in zero-shot inference. While PSALM achieves only 7.9 and 9.6 IoU in the Ego2Exo and Exo2Ego tasks, respectively, our k-NN baseline scores 40.5

and 40.6 IoU while using approximately 10% of the total number of parameters. This highlights the underlying challenges of the cross-view segmentation task and showcases that reformulating the problem as an object mask matching task significantly boosts zero-shot performance.

### 4.4. Ablation study

**O-MaMa architecture.** Tab. 3 details the contribution of each O-MaMa component. Experiment A highlights the benefits of training a simple MLP with our novel Mask Matching Contrastive Loss  $\mathcal{L}_M$ , which aligns cross-view embeddings in a common latent space and improves IoU from 35.2 to 42.2 (Ego2Exo) and 34.9 to 44.7 (Exo2Ego). Second, Experiments A, B and C show that incorporating regional context is only beneficial when we sample adjacent negatives during training, as the hard-mining strategy forces the model to learn more fine-grained discriminative embeddings in nearby candidates with similar context but different mask descriptor. Next, our Ego $\leftrightarrow$ Exo Cross Attention mechanism introduces cross-image content and global information into the object embedding. As Fig. 6 shows, this module incorporates the object features from the other perspective, smoothing the cross-view alignment and improving the final performance. Finally, the joined effect of all our proposed modules specially improves the Loc.E, with relative improvements of +67.5 % (Ego2Exo) and +38.0 % (Exo2Ego), which yields a final gain of +37.2 % Ego2Exo and +42.1 % Exo2Ego IoU. This demonstrates that, while the k-NN baseline is agnostic to the candidate mask location, (it just selects the most similar match),

**Mask Descriptors.** Tab. 4 compares different pooling strategies for obtaining a mask descriptor. The k-NN baseline, which relies solely on object semantic similarity, shows that averaging DINOv2 upsampled features over mask pixels provides the best results (35.2 and 34.9 IoU), as it retains the fine-grained object representation from the dense DINOv2 feature map. This strategy outperforms CLIP-based descriptors (24.5 Ego2Exo and 23.9 Exo2Ego IoU), DINOv2 average pooling over the bounding box (21.8 Ego2Exo and 21.2 Exo2Ego IoU) or DINOv2 mask centroid (25.6 Ego2Exo and 24.1 Exo2Ego IoU). Tab. 4 also

<sup>3</sup>We compute the relative gain% of  $x$  relative to  $y$  as  $100 \cdot (\frac{x-y}{y})$ .

Exp.	$\mathcal{L}_M$	Context	Adj. Neg	C.Attn	Global Union	Ego2Exo			Exo2Ego			Total-IoU $\uparrow$
						IoU $\uparrow$	Loc.E $\downarrow$	Cont.A $\uparrow$	IoU $\uparrow$	Loc.E $\downarrow$	Cont.A $\uparrow$	
Base.	-	-	-	-	-	35.2	0.191	0.455	34.9	0.163	0.423	35.1
A	$\checkmark$	-	-	-	-	42.2	0.074	0.571	44.7	0.112	0.546	43.5
B	$\checkmark$	$\checkmark$	-	-	Concat	42.7	0.069	0.577	44.4	0.116	0.543	43.6
C	$\checkmark$	$\checkmark$	$\checkmark$	-	Concat	46.9	0.079	0.599	45.6	0.107	0.548	46.3
D	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	Weighted Sum.	47.3	0.064	0.611	46.8	0.112	0.543	47.1
E	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	Concat	<b>48.3</b>	<b>0.062</b>	<b>0.621</b>	<b>49.6</b>	<b>0.101</b>	<b>0.576</b>	<b>49.0</b>
Relative Gain % of x with respect to y $\frac{(x-y)}{y}$						+37.2%	+67.5%	+36.5%	+42.1%	+38.0%	+36.2%	+39.6%

Table 3. Ablation study on the O-MaMa proposed modules on the 10% of the validation set.

Method	Geometry	$\mathcal{L}_M$	Ego2Exo			Exo2Ego			Total-IoU $\uparrow$
			IoU $\uparrow$	Loc.E $\downarrow$	Cont.A $\uparrow$	IoU $\uparrow$	Loc.E $\downarrow$	Cont.A $\uparrow$	
Max-Pool(b)-DINOv2 [55]	$\times$	$\times$	5.8	0.306	0.119	17.8	0.216	0.253	11.8
	$\checkmark$	$\times$	6.9	0.302	0.132	20.2	0.210	0.278	13.6
Centroid( $\mathcal{M}$ )-DINOv2 [55]	$\times$	$\times$	25.6	0.202	0.357	24.1	0.178	0.326	24.9
	$\checkmark$	$\times$	26.5	0.190	0.378	26.0	0.172	0.346	26.3
Avg-Pool(b)-DINOv2 [55]	$\times$	$\times$	21.8	0.245	0.324	21.2	0.201	0.291	21.5
	$\checkmark$	$\times$	23.2	0.238	0.345	23.7	0.195	0.314	23.5
Avg-Pool(b)-CLIP [59]	$\times$	$\checkmark$	27.8	0.092	0.426	44.1	<b>0.111</b>	0.537	36.0
	$\times$	$\times$	24.5	0.257	0.325	23.9	0.234	0.301	24.2
	$\checkmark$	$\times$	26.2	0.220	0.359	26.7	0.209	0.335	26.5
Avg-Pool( $\mathcal{M}$ )-DINOv2 [55]	$\times$	$\checkmark$	27.5	0.170	0.379	40.4	0.155	0.477	34.0
	$\times$	$\times$	35.2	0.191	0.455	34.9	0.163	0.423	35.1
	$\checkmark$	$\times$	35.4	0.184	0.467	36.6	0.156	0.440	36.0
	$\times$	$\checkmark$	<b>42.2</b>	<b>0.074</b>	<b>0.571</b>	<b>44.7</b>	0.112	<b>0.546</b>	<b>43.5</b>

Table 4. Ablation study on the mask descriptors and the influence of learning and geometry constraints. We compare the effects of leveraging inferred camera pose constraints or training a simple MLP with our  $\mathcal{L}_M$ , configuration that corresponds to Exp.A in Tab. 3.

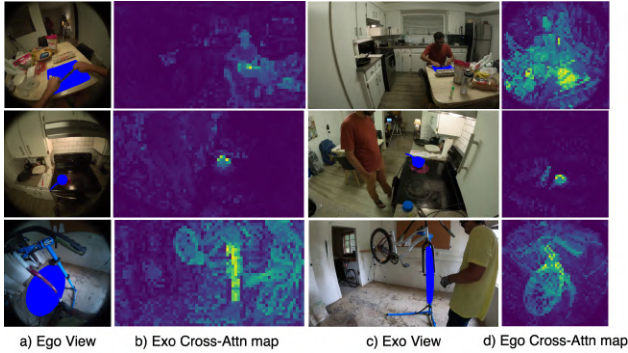


Figure 6. Ego↔Exo Cross-Attention maps

reports that applying geometric constraints yields a minor performance gain (35.2 vs. 35.4 Ego2Exo IoU, and 34.9 vs. 36.6 Exo2Ego IoU when pooling DINOv2 mask features) due to the low success rate of camera pose estimation methods. As Fig. 5 shows, even RoMa [18] struggles with the high viewpoint variance between ego and exo perspectives. This improvement is even less significant when compared to learning view-invariant features with  $\mathcal{L}_M$  (35.2 vs. 42.2 Ego2Exo IoU and 34.9 vs. 44.7 Exo2Ego IoU), highlighting the need to extract stronger visual cues.

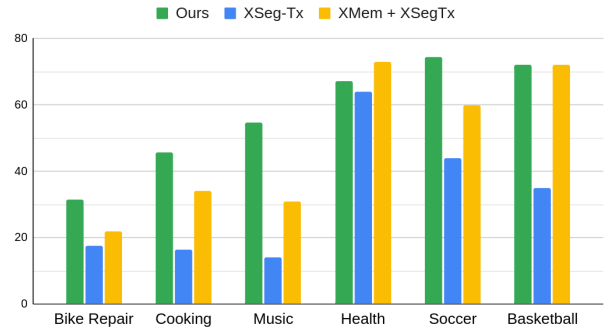


Figure 7. Per-tasks Ego2Exo IoU performance.

**Detailed performance per task and mask size.** Fig. 7 shows the IoU across different scenarios in the Ego2Exo task, where O-MaMa outperforms XMem + XSeg-Tx in most cases, including the challenging *cooking* and *bike repair* activities, which involve cluttered environments and objects of multiple sizes. Fig. 9 analyzes the segmentation performance across the target mask sizes, showing that O-MaMa excels in medium and large-size objects. However, it still struggles with very small objects, as extracting a meaningful mask descriptor remains challenging.



Figure 8. **Exo2Ego Qualitative Results.** We show the source mask in blue and the top 3 target masks in green, yellow and orange.

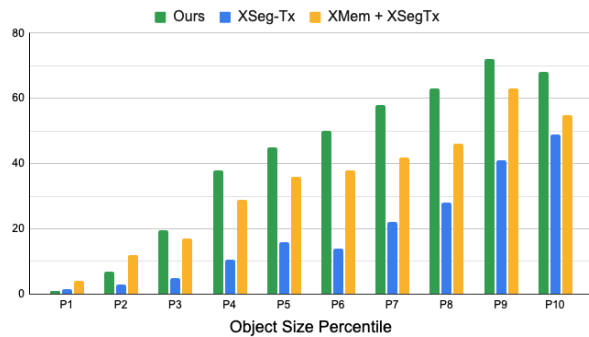


Figure 9. **Ego2Exo IoU performance across different object sizes in the destination view.**

#### 4.5. Qualitative results

We show qualitative examples for the Exo2Ego (Fig. 8) and Ego2Exo (Fig. 10) tasks. The results show that top mask candidates are closely aligned due to their similar context, but our method correctly matches the top-1 mask candidate with the target object. FastSAM’s fine-grained zero-shot capabilities yield high-quality segmentation masks (e.g., the *tire* in Fig. 10 and the *knife* or *bottle* in Fig. 8). However, as a limitation of our approach, they may produce partial segmentations when they capture only a part of the object (e.g., the *saucupan* in Fig. 10). Finally, the total inference time of our approach is 250ms on average, of which 70ms correspond to the FastSAM mask extraction.

### 5. Conclusions

In this work, we address the problem of ego-exo object correspondences, a key step for multi-agent perception. We demonstrate that reformulating cross-view segmentation as an object mask matching problem simplifies the task while improving accuracy under zero-shot conditions. Our Mask Matching Contrastive Loss effectively aligns cross-view embeddings, while DINOv2 pooled mask features preserve fine-grained details. The proposed Hard Negative Adjacent Mining strategy enhances object differentia-



Figure 10. **Ego2Exo Qualitative Results.** For visualization purposes, we show the top 3 masks in green, yellow and orange.

tion, and Ego↔Exo Cross Attention integrates global cross-view context. As a result, O-MaMa achieves state-of-the-art performance on the EgoExo4D Correspondences task while using considerably fewer parameters, obtaining a unified fine-grained segmentation and strong cross-view understanding.

## 6. Acknowledgments

This work was supported by projects PID2021-125209OB-I00 and TED2021-129410B-I00, (MCIN/AEI/10.13039/501100011033 and FEDER/UE and NextGenerationEU/PRTR), and Aragon Government DGA T45-23R.

## References

- [1] Siddhant Bansal, Chetan Arora, and CV Jawahar. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*, pages 657–675. Springer, 2022. 2
- [2] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004. 4
- [3] Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1560–1568, 2022. 2
- [4] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 2
- [5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 1, 2
- [6] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019. 3
- [7] Mai Lee Chang, Greg Trafton, J Malcolm McCurry, and Andrea Lockerd Thomaz. Unfair! perceptions of fairness in human-robot teams. In *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 905–912, 2021. 1
- [8] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020. 2
- [9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 5
- [10] Ho Kei Cheng and Alexander G Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 5
- [11] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021. 2
- [12] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020. 2
- [13] Houston Claire, Mai Lee Chang, Seyun Kim, Daniel Omeiza, Martim Brandao, Min Kyung Lee, and Malte Jung. Fairness and transparency in human-robot interaction. In *17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1244–1246, 2022. 1
- [14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The Epic-Kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2
- [15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2
- [16] Gerard Donahue and Ehsan Elhamifar. Learning to predict activity progress by self-supervised video alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18667–18677, 2024. 2
- [17] Zi-Yi Dou, Xitong Yang, Tushar Nagarajan, Huiyu Wang, Jing Huang, Nanyun Peng, Kris Kitani, and Fu-Jen Chu. Unlocking exocentric video-language data for egocentric video representation learning. *arXiv preprint arXiv:2408.03567*, 2024. 2
- [18] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 2, 5, 7
- [19] Yuqian Fu, Runze Wang, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. Objectrelator: Enabling cross-view object relation understanding in ego-centric and exo-centric videos. *arXiv preprint arXiv:2411.19083*, 2024. 2, 3, 5, 6
- [20] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 2
- [21] Kartik Garg, Sai Shubodh Puligilla, Shishir Kolathaya, Madhava Krishna, and Sourav Garg. Revisit anything: Visual place recognition via image segment retrieval. In *European Conference on Computer Vision*, pages 326–343. Springer, 2024. 4
- [22] Raghav Goyal, Wan-Cyuan Fan, Mennatullah Siam, and Leonid Sigal. Tam-vt: Transformation-aware multi-scale video transformer for segmentation and tracking. *arXiv preprint arXiv:2312.08514*, 2023. 3
- [23] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

- [24] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2, 5, 6
- [25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2
- [26] Matthew S Hutchinson and Vijay N Gadeppally. Video action understanding. *IEEE Access*, 9:134611–134637, 2021. 2
- [27] Helene Intraub. The representation of visual scenes. *Trends in cognitive sciences*, 1(6):217–222, 1997. 4
- [28] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 2
- [29] Marco Karrer, Mohit Agarwal, Mina Kamel, Roland Siegwart, and Margarita Chli. Collaborative 6dof relative pose estimation for two uavs with overlapping fields of view. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6688–6693, 2018. 1
- [30] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023. 2
- [31] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9859–9868, 2020. 2
- [32] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 2
- [33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2, 3, 4, 5
- [34] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2
- [35] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023. 2
- [36] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024. 2
- [37] Liulei Li, Wenguan Wang, Tianfei Zhou, Jianwu Li, and Yi Yang. Unified mask embedding and correspondence learning for self-supervised video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18706–18716, 2023. 3
- [38] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27948–27959, 2024. 2
- [39] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021. 2
- [40] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023. 1
- [41] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 5
- [42] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4463–4472, 2020. 2
- [43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2
- [44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR*, 2019. 5
- [45] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2, 5
- [46] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 45–57, 2023. 2
- [47] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023. 1
- [48] Rohit Mohan and Abhinav Valada. Efficienttps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 129(5):1551–1579, 2021. 2
- [49] Lorenzo Mur-Labadia, Jose J Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5238–5249, 2023. 2

- [50] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Jose J Guerrero. Bayesian deep learning for affordance segmentation in images. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6981–6987, 2023. 2
- [51] Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Jose J Guerrero, Giovanni Maria Farinella, and Antonino Furnari. Aff-tention! affordances and attention models for short-term object interaction anticipation. In *European Conference on Computer Vision*, pages 167–184. Springer, 2024. 2
- [52] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 163–172, 2020. 2
- [53] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3
- [54] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5
- [55] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4, 5, 7
- [56] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 2, 3, 5
- [57] Camillo Quattrocchi, Antonino Furnari, Daniele Di Mauro, Mario Valerio Guffrida, and Giovanni Maria Farinella. Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs. In *European Conference on Computer Vision*, pages 253–270. Springer, 2024. 2
- [58] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5213–5224, 2023. 2
- [59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7
- [60] H. Rahmani and A. Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *CVPR*, 2016. 2
- [61] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1939–1946, 2013. 3
- [62] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2
- [63] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2, 5
- [64] Patrik Schmuck and Margarita Chli. Ccm-slam: Robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams. *Journal of Field Robotics*, 36(4):763–781, 2019. 1
- [65] Xi Shen, Alexei A Efros, Armand Joulin, and Mathieu Aubry. Learning co-segmentation by segment swapping for retrieval and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5082–5092, 2022. 3, 5
- [66] Michal Shlapentokh-Rothman, Ansel Blume, Yao Xiao, Yuqun Wu, Sethuraman TV, Heyi Tao, Jae Yong Lee, Wilfredo Torres, Yu-Xiong Wang, and Derek Hoiem. Region-based representations revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17107–17116, 2024. 4
- [67] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2
- [68] Tatsunori Tanai, Sudipta N Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4246–4255, 2016. 3
- [69] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *Proceedings of the IEEE international conference on computer vision*, pages 4481–4490, 2017. 2
- [70] Tram Thi Minh Tran, Shane Brown, Oliver Weidlich, Mark Billinghurst, and Callum Parker. Wearable augmented reality: Research trends and future directions from three major venues. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4782–4793, 2023. 1
- [71] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3899–3908, 2016. 2
- [72] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 5
- [73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [74] Daiwat Amit Vyas and Dvijesh Bhatt. Augmented reality (ar) applications: A survey on current trends, challenges, & future scope. *International Journal of Advanced Research in Computer Science*, 8(5), 2017. 1

- [75] Mengmeng Wang, Teli Ma, Shuo Xin, Xiaojun Hou, Jiazheng Xing, Guang Dai, Jingdong Wang, and Yong Liu. Visual object tracking across diverse data modalities: A review. *arXiv preprint arXiv:2412.09991*, 2024. 2
- [76] Sanghyun Woo, Dahun Kim, Joon-Young Lee, and In So Kweon. Learning to associate every segment for video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2705–2714, 2021. 2
- [77] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 1, 2
- [78] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16111–16121, 2024. 2
- [79] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018. 3
- [80] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023. 2
- [81] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 931–940, 2019. 2
- [82] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14*, pages 467–483. Springer, 2016. 2
- [83] H. Yu, M. Cai, Y. Liu, and F. Lu. Joint attention learning for first and third person video co-analysis. *ACM MM*, 2019. 2
- [84] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023. 2
- [85] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12): 14679–14694, 2023. 5, 6
- [86] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 2, 3, 5, 6
- [87] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 2, 3, 4, 5
- [88] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023. 5
- [89] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15116–15127, 2023. 2